HGG Advances

ARTICLE

Using GWAS summary data to impute traits for genotyped individuals

Jingchen Ren,[1,2] Zhaotong Lin,[2] Ruoyu He,[1,2] Xiaotong Shen,[1] and Wei Pan[2,3,*]

Summary

Genome-wide association study (GWAS) summary data have become extremely useful in daily routine data analysis, largely facilitating new methods development and new applications. However, a severe limitation with the current use of GWAS summary data is its exclusive restriction to only linear single nucleotide polymorphism (SNP)-trait association analyses. To further expand the use of GWAS summary data, along with a large sample of individual-level genotypes, we propose a nonparametric method for large-scale imputation of the genetic component of the trait for the given genotypes. The imputed individual-level trait values, along with the individual-level genotypes, make it possible to conduct any analysis as with individual-level GWAS data, including nonlinear SNP-trait associations and predictions. We use the UK Biobank data to highlight the usefulness and effectiveness of the proposed method in three applications that currently cannot be done with only GWAS summary data (for SNP-trait associations): marginal SNP-trait association analysis under non-additive genetic models, detection of SNP-SNP interactions, and genetic prediction of a trait using a nonlinear model of SNPs.

Introduction

Genome-wide association studies (GWASs) have been quite successful in identifying genetic variants, mainly single nucleotide polymorphisms (SNPs), associated with complex traits and common diseases.[1,2] In particular, the increasing availability of GWAS summary data (for SNP-trait association as usually used and assumed throughout this paper) has been playing an indispensable role in largely facilitating the development of new methods for new applications and secondary data analyses,[3] such as construction of polygenic risk scores (PRSs) for trait prediction,[4] fine mapping,[5] heritability estimation,[6–8] genetic correlation analysis,[9,10] causal inference with Mendelian randomization,[11,12] and transcriptome-wide association studies,[13,14] just to name a few; see Pasaniuc and Price[15] for a recent review. Nevertheless, since marginal SNP-trait association estimates in GWAS summary data measure only linear relationships between the SNPs and trait, the current use of GWAS summary data is limited to exploiting only linear SNP-trait associations; it is unknown how to use GWAS summary data for nonlinear SNP-trait association analysis. For example, given a GWAS summary dataset (and a reference panel of individual-level genotypes), it seems impossible to detect SNP-SNP interactions[16,17] or to build a PRS model accounting for possibly nonlinear and epistatic SNP effects by taking advantage of many emerging powerful nonlinear machine learning models such as random forests (RFs)[18–22] and deep learning.[23,24] These nonlinear SNP-trait association and prediction analyses are expected to shed more light on the genetic architecture of complex traits, deepen mechanistic understanding of common diseases, and thus advance translational applications of genetics.

Here we point out that it is possible for nonlinear modeling and analyses based on a GWAS summary dataset *and* a sample of individual-level genotypes. More generally, with a GWAS summary dataset of a trait, we can impute the trait values for a large sample of genotypes, which can be useful if the trait is not available, either unmeasured or difficult to measure (e.g., status of a late-onset disease), in a biobank. We propose a nonparametric method for large-scale imputation of the genetic component of a trait for each of the individuals with (genome-wide) genotypic data. With the individual-level genotypes and imputed trait values, one can conduct any (linear or nonlinear) GWAS analysis as with individual-level data. We use the UK Biobank data to show that, with GWAS summary data and individual-level genotypes (that may or may not be different from those of the summary data), using the imputed trait values for subsequent analyses led to results quite similar to that obtained from using the observed/true trait values. In particular, we showcase three applications with trait imputation: given a GWAS summary dataset, we conducted marginal SNP-trait association analysis under a non-additive genetic model for SNPs,[25,26] detecting SNP-SNP interactions, and predicting a complex trait using a nonlinear model (i.e., RFs) for a sample of individuals with only genotypic data. Since any existing PRS method can be applied to impute the trait values for a sample of genotypes, one may wonder how our proposed method compares with existing PRS methods. In short, because existing PRS methods for GWAS summary data are all based on some assumed linear models, they will not be suitable for subsequent association

[1]School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA; [2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA
[3]Lead contact
*Correspondence: panxx014@umn.edu
https://doi.org/10.1016/j.xhgg.2023.100197.

analyses. We used a state-of-the-art PRS method[27] as a representative to confirm the point in the above three applications. Given the increasing availability of GWAS summary data of various traits and of large-scale biobanks with genotypic but not phenotypic data for some traits of interest, we expect that the proposed method will further expand the use of GWAS summary data in many new applications, especially for nonlinear or/and integrative analysis of GWAS summary data and biobank data.

Material and methods

Overview

For an individual with genome-wide genotype (score) vector x and a quantitative trait y, we assume a genetic model:

$$y = E(y|x) + \epsilon = g(x) + \epsilon, \qquad \text{(Equation 1)}$$

where $E(y|x) = g(x)$ is the genetic (or genetically regulated) component of the trait that is unknown, and ε captures all other genetically independent environmental effects and white noises. Note that the functional form of $g(x)$ is unspecified; in particular, it may not be linear in x, thus allowing and accounting for nonlinear and epistatic effects of the SNPs.

Suppose we have a GWAS summary dataset $\{(\widehat{\beta}_j^*, \sigma_j^*) : j = 1, \ldots, p\}$ for p SNPs based on an individual-level GWAS dataset (X^*, Y^*), called training data. As usual, the individual-level data are not available. Each $\widehat{\beta}_j^*$ is the ordinary least-squares estimator (OLSE) of the marginal association between SNP j and the trait, and $\sigma_j^* = \mathrm{SE}(\widehat{\beta}_j^*)$ is the standard error (SE). Now given an individual-level genotypic dataset X for a sample of (approximately) unrelated individuals from the same population, called the test data, we would like to recover (the genetic component of) the trait for each individual in the test sample. We developed a nonparametric method for this purpose without specifying the functional form of $g(.)$; in contrast, all existing PRS methods for summary data are based on a parametric linear model for $g(.)$. The main idea is that, if we had Y, we could estimate the marginal association as $\widehat{\beta}_j$ for each SNP j; under the assumption that both the training and test data come from the same population, with large sample sizes n_1 and n_2, we have $\widehat{\beta}^* \approx \widehat{\beta}$, which can be used to formulate a least-squares (LS) problem to estimate the genetic components of Y. We loosely call the procedure imputing or recovering (the genetic components of) Y.

After obtaining $\widehat{Y}$, we can conveniently treat $(X, \widehat{Y})$ as an individual-level GWAS dataset for subsequent analyses, including assessing nonlinear SNP-trait associations that cannot be accomplished with the original GWAS summary data alone, such as detecting SNP-trait associations under a non-additive model and detecting SNP-SNP interactions as to be illustrated next.

We took a random sample of $n_1 = 178,175$ individuals from the UK Biobank and used their genotypes and trait high-density lipoprotein cholesterol (HDL) as the training data (to obtain and then use the GWAS summary data) and those for the other $n_2 = 178,176$ as the test data. We imputed the trait values for the test data and compared the performance of using the observed trait values with that of using imputed ones in various tasks.

The LS-imputation method

Suppose we have a GWAS summary dataset $\{(\widehat{\beta}_j^*, \sigma_j^*) : j = 1, \ldots, p\}$ for p SNPs derived from an individual-level GWAS dataset of n_1 individuals, (X^*, Y^*), with $X^* = (X_{\cdot 1}^*, X_{\cdot 2}^*, \ldots, X_{\cdot p}^*)$, $X_j^* = (X_{1j}^*, X_{2j}^*, \ldots, X_{n_1 j}^*)'$, and $Y^* = (Y_1^*, Y_2^*, \ldots, Y_{n_1}^*)'$, called training data. As usual, only the GWAS summary data, not the individual-level training data, are available; we use the individual-level data here only for the purpose of notation or illustration. We assume throughout that each SNP is centered to have sample mean 0; although not required, for simplicity of notation, we also assume that each SNP is scaled to have sample variance 1. Then the OLSE $\widehat{\beta}^* = (\widehat{\beta}_1^*, \ldots, \widehat{\beta}_p^*)'$ of the marginal associations between the SNPs and the trait is

$$\widehat{\beta^*} = \frac{1}{n_1 - 1} X^{*\prime} Y^*, \text{ or equivalently,}$$

$$\widehat{\beta_j^*} = \left(X_j^{*\prime} X_j^*\right)^{-1} X_j^{*\prime} Y^* = \frac{1}{n_1 - 1} X_j^{*\prime} Y^*.$$

Given a new test dataset of n_2 individuals with only genotypic data as an $n_2 \times p$ SNP matrix X, we would like to impute the corresponding (genetic components) of the trait vector Y. If Y were available, we would estimate the marginal association effects as $\widehat{\beta} = \frac{1}{n_2 - 1} X'Y$. Since both $\widehat{\beta}$ and $\widehat{\beta}^*$ are consistently estimating the same true (and unknown) marginal association parameter β, if the sample sizes are not too small, they should be close to each other. Hence, to impute Y, we consider the following LS problem:

$$\begin{aligned} \widehat{Y} &= \arg\min_Y \left\| \widehat{\beta}^* - \frac{1}{n_2 - 1} X'Y \right\|^2 \\ &= (n_2 - 1)(XX')^+ X \widehat{\beta}^* \\ &= (n_2 - 1) X'^+ \widehat{\beta}^*. \end{aligned} \qquad \text{(Equation 2)}$$

In the above OLSE, due to centering of each SNP (i.e., column of X) at sample mean 0, X is not of full rank, so the Moore-Penrose generalized inverse is used (while other generalized inverse can be equally used but will not be pursued here). It is easy to see that the solution to the above LS problem is not unique: given any solution $\widehat{Y}$, $\widehat{Y} + c$ for any constant c is a solution too, due to $X'1c = 0$. Alternatively, some regularization via a small ridge penalty can be added into the objective function to obtain a unique solution $\widehat{Y}(\lambda) = (n_2 - 1)(XX' + \lambda I)^{-1} X \widehat{\beta}^*$; it turns out that $\widehat{Y}(\lambda)$ with a small $\lambda > 0$ is similar to $\widehat{Y}$ because the Moore-Penrose generalized inverse can be expressed as $X'^+ = \lim_{\lambda \to 0^+} (XX' + \lambda I)^{-1} X$.

As detailed in the supplemental information, we compared several implementations and found out that using $\widehat{Y}(\lambda)$ with $\lambda = 10^{-6}$ was computationally both fast and stable, and thus it was chosen as the default to be used in the paper. Furthermore, if n_2 is too large, we may not be able to invert the corresponding matrix within a reasonable amount of computing time or memory; by default, we will divide the test dataset into smaller batches of size m. As to be discussed next, we will try a few m values and choose the one that gives marginal (additive) association results similar to that from the training data. Note that we require $p > m$ (or $p > n_2$ if no batch is used), while preferring to have both n_1 and p as large as possible.

Extensions

As discussed exclusively in the supplemental information, if we have the estimated intercept in the marginal regression model for each SNP-trait pair, we can apply the same method but do not need to center the SNP matrix X and X is of full rank, leading to $(XX')^+ = (XX')^{-1}$, thus some simpler results and interpretations.

Furthermore, instead of using the OLSE, we can use the weighted least-squares (WLS) method with the weights inversely proportional to the variances of the elements of $\widehat{\beta}^*$; but as shown in the supplemental information, its results were similar to those from the OLSE. Finally, as shown in the supplemental information, our method can be extended to binary traits.

Key features

To gain some intuitive understanding of the LS-imputation method, we consider a simpler scenario with the intercept known in the marginal regression model for each SNP-trait pair, for which no centering on X or X^* is needed and we have $(XX')^+ = (XX')^{-1}$. The LS-imputed trait is

$$\widehat{Y} = \frac{n_2 - 1}{n_1 - 1}(XX')^{-1}XX^{*\prime}Y^*,$$

a linear combination of the trait values Y^* in the training data. $XX' = (S_{ij})$ and $XX^{*\prime} = (S_{ij^*})$ can be regarded as some genotypic similarity matrices for the individuals in the test data and those between the test and training data, respectively. In other words, the imputed trait value for a test individual is some weighted average of the trait values in the training data, where the weights are determined by the genotypic similarities of this test individual with others in both the training and test data. Consequently, the imputed trait $\widehat{Y}$ may contain nonlinear genetic information embedded in the observed trait Y^*. To make this clear, let us consider a special case: if $X = X^*$, we have $\widehat{Y} = Y^*$; that is, if we have the genotypes of the training data, we would perfectly recover their trait values. Other more general cases with overlapping individuals (or equal genotypes) are discussed in supplemental information S1.1.3. More generally, to be concrete, let us consider $n_2 = 2$ individuals in the test data. It is easy to derive

$$\widehat{Y} = \begin{pmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \end{pmatrix} = \frac{1}{(n_1 - 1)(S_{11}S_{22} - S_{12}^2)} \begin{bmatrix} \sum_{j=1}^{n_1}(S_{22}S_{1j^*} - S_{12}S_{2j^*})Y_j^* \\ \sum_{j=1}^{n_1}(S_{11}S_{2j^*} - S_{12}S_{1j^*})Y_j^* \end{bmatrix}.$$

It not only confirms that an imputed value is a linear combination of the trait values in the training data, but it also illustrates a distinct feature of the method: an imputed trait not only depends on its own genotype but also other genotypes in the test data; that is, there is information borrowing across similar individuals in the test data. On the other hand, if the individuals are not similar at all with $S_{12} = 0$, we have $\widehat{Y}_i = \sum_{j=1}^{n_1} S_{1j^*}Y_j^*/[(n_1 - 1)S_{ii}]$ for $i = 1, 2$, suggesting that an imputed value only depends on its genotypic similarities to those in the training data.

Statistical properties of LS-imputation

Since XX' plays a key role in our method, we show in the supplemental information that XX'/p behaves like the centering matrix $C_{n_2} = I - 11'/n_2$ for a large p, which will offer some insights on the properties of the LS-imputation method. Note that we use 1 to also represent a vector with all elements' 1's.

We first consider a special case: when we have the genotypes of the training data, the LS-imputed Y would (asymptotically) recover the centered trait values in the training data. Specifically, if $X = X^*$, we have

$$\widehat{Y} = (X^*X^{*\prime})^+X^*X^{*\prime}Y^* = (X^*X^{*\prime})^+X^*X^{*\prime}(C_{n_2}Y^*),$$

due to centered $X^* = C_{n_2}X^*$, suggesting that the imputed $\widehat{Y}$ is a linear transformation of the centered Y^*. If we have independent individuals and independent SNPs (or locally dependent SNPs satisfying the central limit theorem[28]), as $p \to \infty$, we have $XX'/p \xrightarrow{P} C_{n_2}$, as shown in the supplemental information. Moreover, we have $C_{n_2}^+ = C_{n_2} = C_{n_2}^2$ and thus

$$\widehat{Y} \xrightarrow{P} C_{n_2}{}^+C_{n_2}Y^* = C_{n_2}Y^*.$$

That is, the imputed trait values would tend to the (centered) trait values in the training data, which obviously contain information about possible *nonlinear* SNP-trait associations, even though *linear* SNP-trait associations $\widehat{\beta}^*$ are used for trait imputation. This result may sound trivial, but it is unique to our method and offers some insights about our method; we are not aware of any existing PRS method possessing this property.

More generally, we have

$$\widehat{Y} = \frac{n_2 - 1}{n_1 - 1}(XX')^+XX^{*\prime}Y^* \approx \frac{n_2 - 1}{n_1 - 1}(XX^{*\prime}/p)C_{n_1}Y^*,$$

suggesting that the imputed trait values for the test data are linear combinations of the (centered) trait values in the training data, where the weights are the genotypic similarities between the test and training data as measured by $XX^{*\prime}/p$. In other words, an imputed trait value is a linear combination of the (centered) trait values in the training data, implying its possibly recovering/containing information about linear or nonlinear SNP-trait associations. Furthermore, if n_1 is large (as usual), we have $X^{*\prime}Y^* = X^{*\prime}g(X^*) + X^{*\prime}\epsilon \approx X^{*\prime}g(X^*)$, where $g(X^*) = (g(X_1^*), g(X_2^*), \ldots, g(X_{n_1}^*))'$ is the vector of the genetic components of the individuals in X^* with X_i^* being the genotype vector for individual i. Thus

$$\widehat{Y} \approx \frac{n_2 - 1}{n_1 - 1}(XX^{*\prime}/p)C_{n_1}g(X^*),$$

indicating that the imputed trait values are related to their genetic components.

The variance of $\widehat{Y}$ can be calculated as

$$\mathrm{Var}(\widehat{Y}) := \mathrm{Var}(\widehat{Y}|X, X^*) = (n_2 - 1)^2(XX')^+X\mathrm{Var}(\widehat{\beta}^*)X'(XX')^+, \quad \text{(Equation 3)}$$

suggesting that in general the elements of $\widehat{Y}$ are correlated (and have unequal variances). However, since it will be difficult to invert the large matrix XX' (while being complicated to deal with the elements that are not independent and identically distributed [iid] of $\widehat{Y}$ in subsequent analyses), a simple way we take is to (incorrectly) treat the elements of $\widehat{Y}$ as independent (with an equal variance) in subsequent analyses, which, as to be shown, may lead to slightly over- or under-estimating SEs; but with a suitable choice of batch size, the problem was largely negligible.

To gain more insights, we consider a special case with elements of X being iid standard normal and X being of full rank (i.e., no centering of each column of X), under which we have

$$\mathrm{Var}(\widehat{Y}_j) = (n_2 - 1)^2(XX')_{jj}^{-1}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\tau^2,$$

where $(XX')_{jj}^{-1}$ is the j th diagonal element of matrix $(XX')^{-1}$, and $\tau^2 = \mathrm{Var}(X_{ij}^*Y_i^*)$. We show the following in the supplemental information. First, if n_2 is small (and fixed), by the strong law of large numbers, we have $\mathrm{Var}(\widehat{Y}_j) \approx n_2\tau^2/p$. Second, if $n_2 < p$ is large with $n_2/p \to c \in (0, 1)$ as both n_2 and p go to ∞, using random matrix theory,[29] we have

$$\mathrm{Var}(\widehat{Y}_j) = n_2O(1)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\tau^2.$$

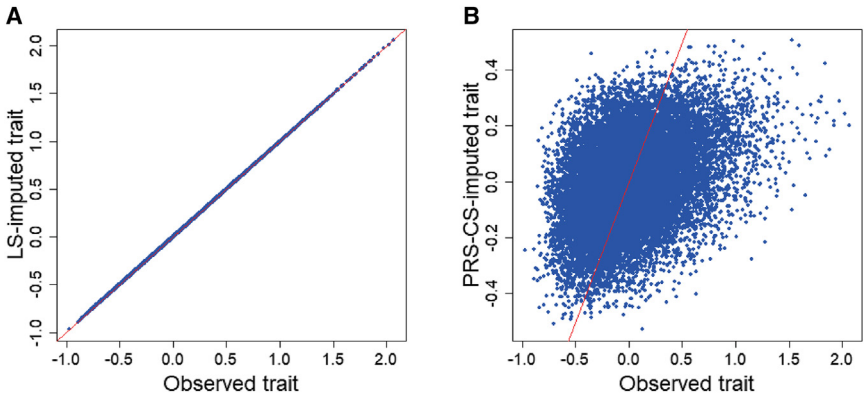

Figure 1. Comparison of the observed and imputed (HDL) trait values on one batch of the training data: (A) by our LS-imputation method; (B) by PRS-CS

As discussed in more detail in the supplemental information, these results suggest that we should use a small n_2 to obtain small variances for imputed $\widehat{Y}$. Finally and more generally, although the SNPs are not iid normal, the above random matrix theory can be extended to dependent non-normal distributions[29] and serve as a good approximation for SNP data.[30]

A practical question is, given a large sample of genotypes, from both computational and statistical considerations, whether and how we should divide it into smaller batches of size m so that we can impute Y for each batch separately. Based on the above analyses, on one hand, a smaller m leads to smaller variances of the elements of $\widehat{Y}$; on the other hand, because of imputed $\widehat{Y}$ for each batch being centered at mean 0, there would be information loss between any two batches because they may be no longer comparable. Considering two batches with true $Y_{(1)}$ and $Y_{(2)}$ with the former stochastically larger than the latter, since both $\widehat{Y}_{(1)}$ and $\widehat{Y}_{(2)}$ will be centered at mean 0, we lose the information that the elements of $Y_{(1)}$ tend to be larger than those of $Y_{(2)}$. Note that, with a large m, by the law of large numbers, the means of the Y's in the batches will be close, so the centering effects of imputing on each batch will be small. Hence, we currently suggest using a relatively large batch size m that is computationally feasible and gives marginal analysis results similar to those of the training data.

UKB data

We used the UK Biobank (UKB) data[31] to illustrate the application of our proposed method. Specifically, we considered trait HDL and 356,351 individuals of White British ancestry with no missing value of HDL. Starting with the imputed genotypic data, we filtered out the SNPs each with minor allele frequency less than 0.05, with missing values larger than 10%, or failing the Hardy-Weinberg equilibrium exact test with p value less than 0.001. Furthermore, we pruned out SNPs in high linkage disequilibrium (LD) with a window size of 50, a step size of 1, and an r^2 threshold of 0.8. After these steps, we ended up with 715,783 SNPs.

We then randomly split the data into two parts, one as the training set X^* of dimension $178,175 \times 715,783$ and the other as the test set X of $178,176 \times 715,783$. Both X^* and X contained some missing values, each of which was replaced/imputed with the mean of the observed values of the corresponding SNP. Both X^* and X were centered (so that each SNP had mean 0). We used the training data to calculate the estimated marginal effects $\widehat{\beta}^*$, their SEs, and the p values. Our primary goal was to use the (training set-based) GWAS summary data and the individual-level genotypic test data X to impute Y; we compared the performance of using imputed $\widehat{Y}$ with that of using the observed Y.

Implementation details

Due to computational and data storage limitations, we could not use all the 715,783 SNPs to apply our method. Instead, we selected 50,000 SNPs in subsequent analyses: in the training data, we had 67,036 SNPs with p values less than 0.05, from which we randomly selected 50,000 SNPs and used them throughout (unless specified otherwise).

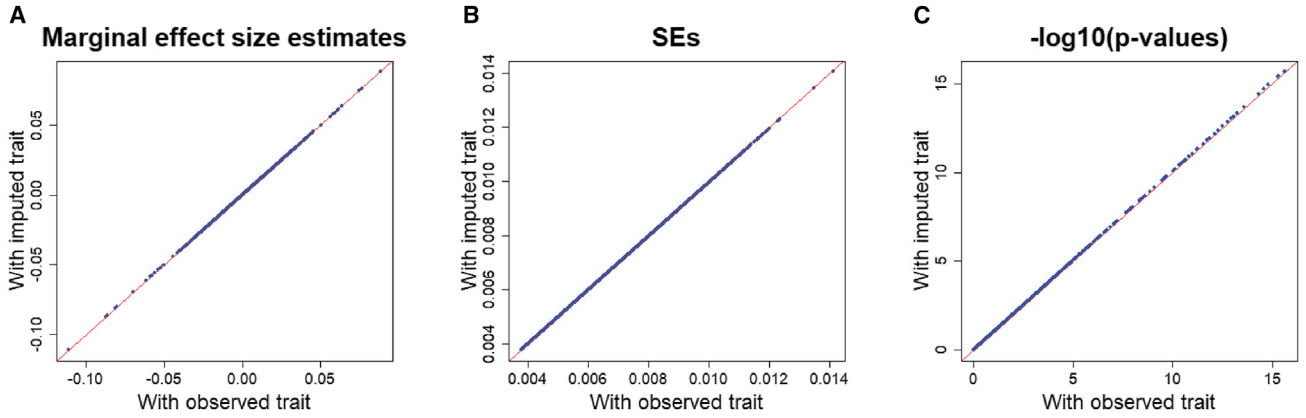

Figure 2. Comparison of the estimated marginal effect sizes (A), their SEs (B), and –log10(p values) (C) calculated with the observed and LS-recovered (HDL) trait values for one batch of the training data

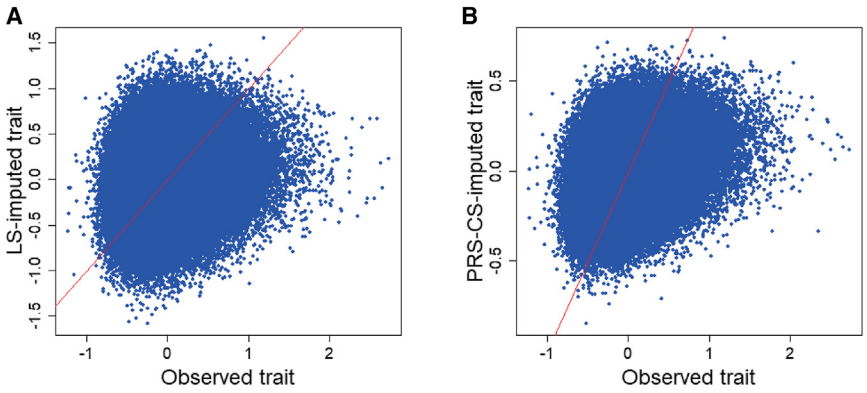

Figure 3. Comparison of the observed and imputed (HDL) trait values on the test data: (A) by our LS-imputation method; (B) by PRS-CS

Similarly, there may not be enough computer memory to hold data for all n_2 individuals in the test data, or it would take too long to invert the corresponding matrices; furthermore, as analyzed earlier, to achieve a better bias-variance trade-off, it may be better to impute for a smaller set of individuals at one time. Accordingly, in our example, we split the test data into nine batches of almost equal sizes, eight with sample size $m = 20,000$ and one with $m = 18,176$. We applied the LS-imputation method to each batch separately and then pooled the imputed trait values across the batches together. In implementing our method, we used `linalg.inv` function in Python package numpy to invert a matrix (with all the parameters in the function set to their default values).

Based on the previous theoretical analysis, we would recommend using p and n_1 as large as possible. Even if it is computationally feasible to deal with a large n_2, we recommend using a smaller $m < p$.

PRS-CS

PRS-CS is a PRS method based on a high-dimensional Bayesian linear regression model:

$$Y = X\beta + \epsilon,$$

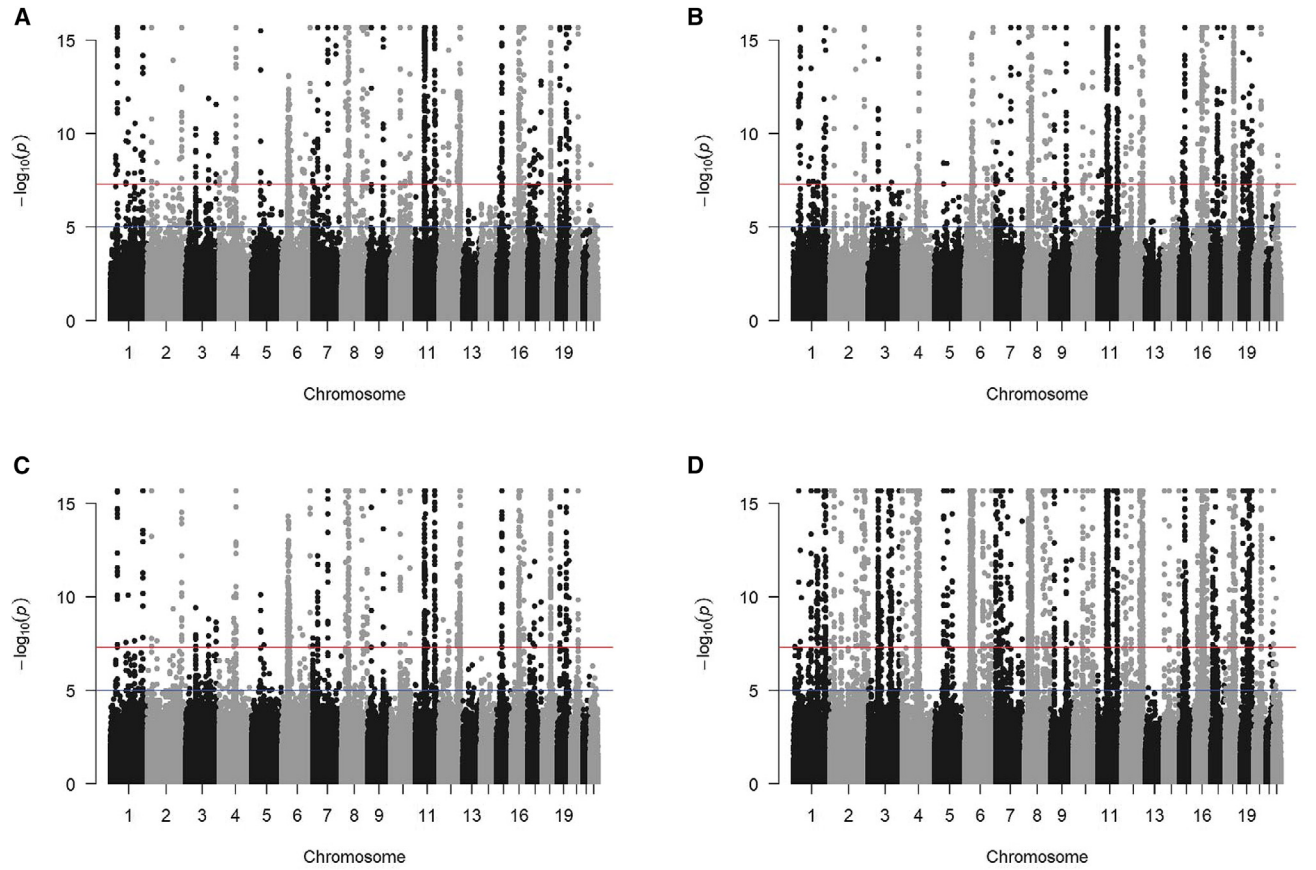

Figure 4. Manhattan plots under the additive model: (A) observed/true trait (HDL) in the training data; (B) observed trait in the test data; (C) LS-imputed trait in the test data; (D) PRS-CS-imputed trait in the test data

The horizontal red and blue lines correspond to the usual and suggestive genome-wide significance levels of 5×10^{-8} and 10^{-5} respectively.

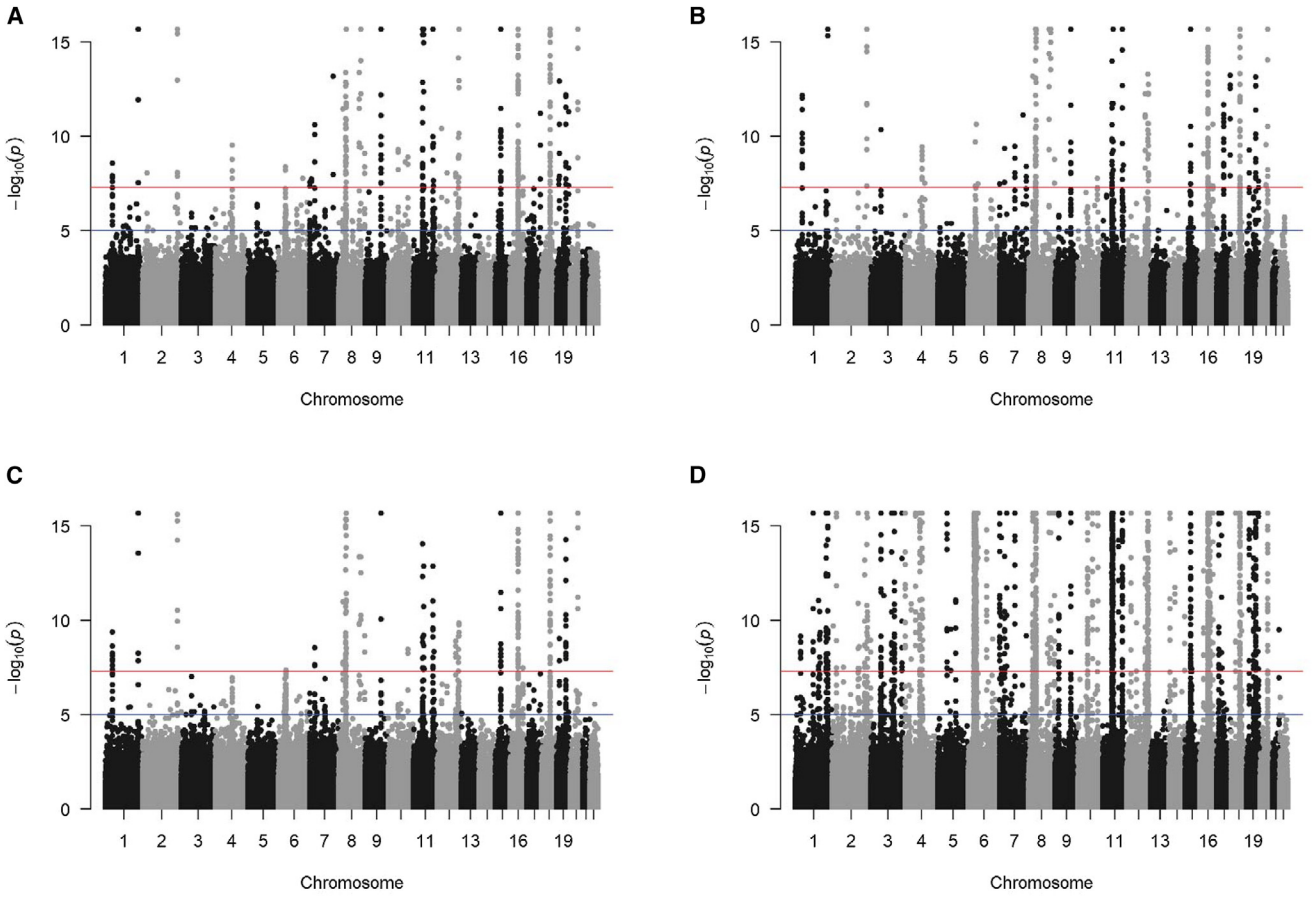

Figure 5. Manhattan plots under the recessive model: (A) using the observed/true trait (HDL) in the training data; (B) observed trait in the test data; (C) LS-imputed trait in the test data; (D) PRS-CS-imputed trait in the test data
The horizontal red and blue lines correspond to the usual and suggestive genome-wide significance levels of 5×10^{-8} and 10^{-5} respectively.

where a continuous shrinkage (CS) prior is put on β to improve the prediction accuracy.[27] It has been shown to be a top performer among the existing PRS methods.[4,32,33] As for LS-imputation, we first apply PRS-CS to a training GWAS summary dataset to estimate the parameters in the model and then use the fitted model to predict/impute Y for any given test genotypes X.

To implement PRS-CS, we used the PRS-CS-auto version (without a separate validation dataset for tuning parameter selection) in the software provided by the original authors of the PRS-CS paper (https://github.com/getian107/PRScs). The software requires GWAS summary statistics and a reference panel (to estimate the LD structure) as the input. In our analysis, we chose the UK Biobank EUR data provided by the software as the reference panel. All the parameters in the software just used the default values. We applied PRS-CS(-auto) to the pre-processed UKB GWAS data with all 715,783 SNPs; at the end, it retained 120,634 SNPs to calculate a polygenic risk score for HDL in the main results.

Statistical analysis

Statistical analyses, including that for marginal SNP-trait association and SNP-SNP interactions, were conducted in R. RF models were fitted using the RandomForestRegressor function in Python package sklearn.

Results

Recovering the trait values in the training data

First we would like to confirm that our LS-imputation method can (almost) perfectly recover the trait values if the same genotypic data $X = X^*$ as in the training data are used. Figure 1 compares the observed and imputed trait values for a batch by our LS-imputation method and PRS-CS method. We could see that the LS-imputation method, but not PRS-CS, could almost perfectly recover the trait values in the training data.

Figure 2 shows the scatterplots of the estimated marginal effect sizes, SEs of the estimated marginal effects, and -log10(p values) calculated with the observed and LS-recovered trait (HDL) values on one random batch ($n = 20,000$) of the training set respectively; the corresponding (Pearson) correlations are all > 0.999. We could see that our method performed extremely well on the estimation of the marginal effects, SEs, and the p values for the SNPs because it gave the results almost exactly the same as those from using the observed trait values.

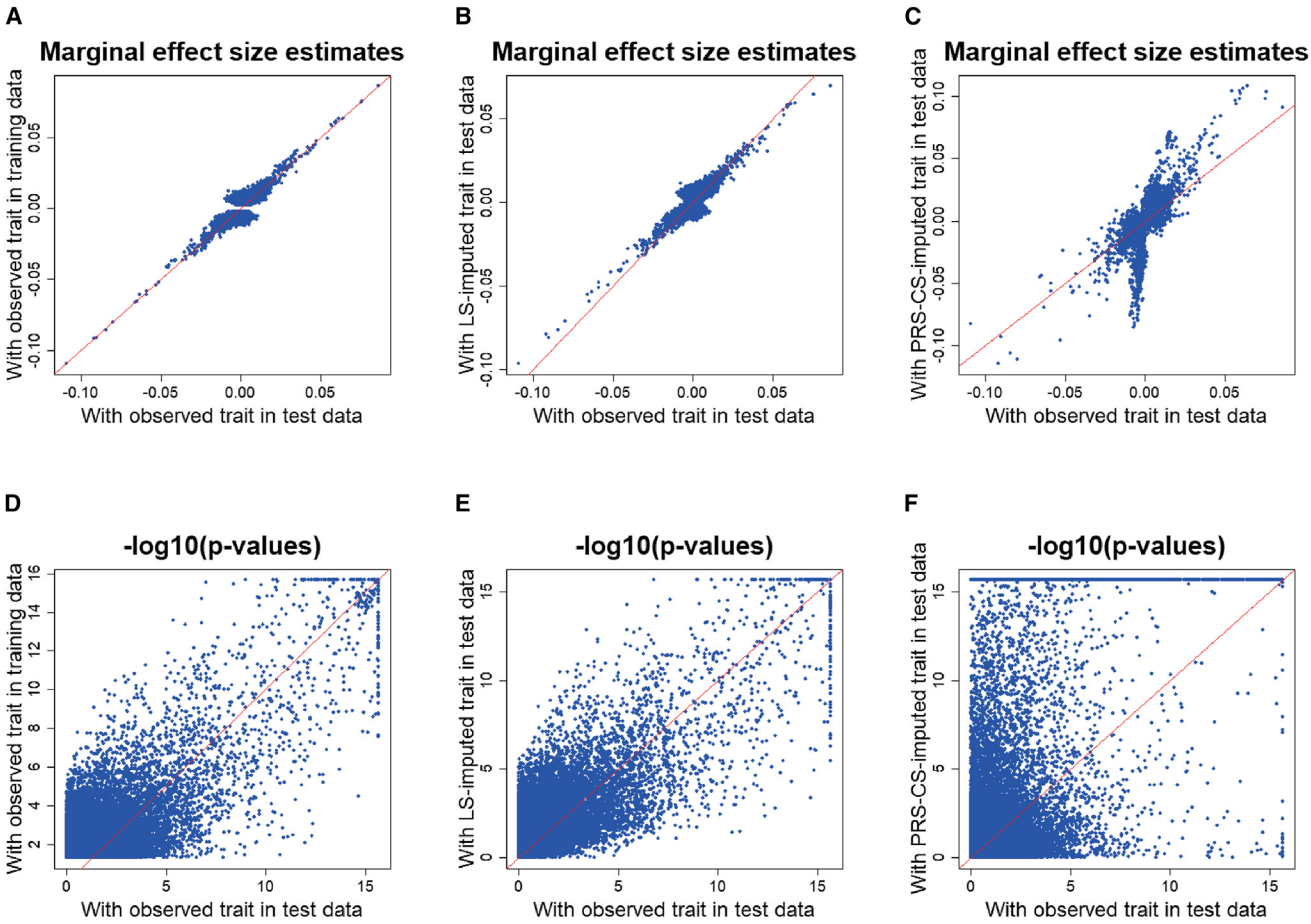

Figure 6. Comparison of the estimated marginal effect sizes (A–C) and –log10(p values) (D–F) under the additive model: calculated (A and D) with the observed (HDL) trait (and genotypes) in the training and test data respectively; (B and E) with the observed (HDL) trait in the test data and LS-imputed trait in the test data; and (C and F) with the observed (HDL) trait in the test data and PRS-CS-imputed trait in the test data

Imputing the trait values in the test data

Figure 3 compares the observed and imputed trait values on the test data. For comparison, we also show the results from PRS-CS. The corresponding correlations between the observed and imputed trait values were 0.177 and 0.279 for the LS and PRS-CS methods respectively. Note that throughout this paper we did not adjust for any covariates; as expected and shown in the supplemental information, if we adjusted for sex and age, the correlations between the observed and LS- or PRS-CS-imputed trait values for the test data were slightly larger at 0.204 and 0.313 respectively. The linear model-based PRS-CS did better in trait prediction than the nonparametric LS-imputation method, perhaps because linear/additive effects of SNPs dominated the heritability.[34,35] Nevertheless, as to be shown later, the LS-imputation method performed much better in subsequent association analyses, including nonlinear analyses.

Marginal associations under various genetic models for SNPs

Next we demonstrate a main advantage of our method: the imputed trait values for the (test) genotypic data can be used to detect nonlinear effects of SNPs; in contrast, we cannot do so based on the original/training GWAS summary data or imputed trait values derived from a standard PRS method. Specifically, we will consider another two genetic models in addition to the additive model that is used by default in GWAS as in any generated GWAS summary data. For better visualization throughout this paper, we truncated any p value $< 2.2\times10^{-16}$ at 2.2×10^{-16}.

We first consider the usual additive model with the results shown in Figure 4: panels A and B are based on the observed trait (HDL) values in the training and test data, respectively, while panels C and D are the imputed trait values based on our LS-imputation method and PRS-CS, respectively, for the test data. We could see that the distributions of the significant SNPs identified with the observed trait on the training and test data look quite similar, and more importantly, they are similar to that identified with our LS-imputed trait, though our method gave slightly more conservative results with less and fewer significant SNPs. In contrast, PRS-CS identified way too many significant SNPs; in fact, any SNPs used in the PRS-CS (or any other PRS) model and those in LD with them, by definition, will be significant if the sample size is large enough.

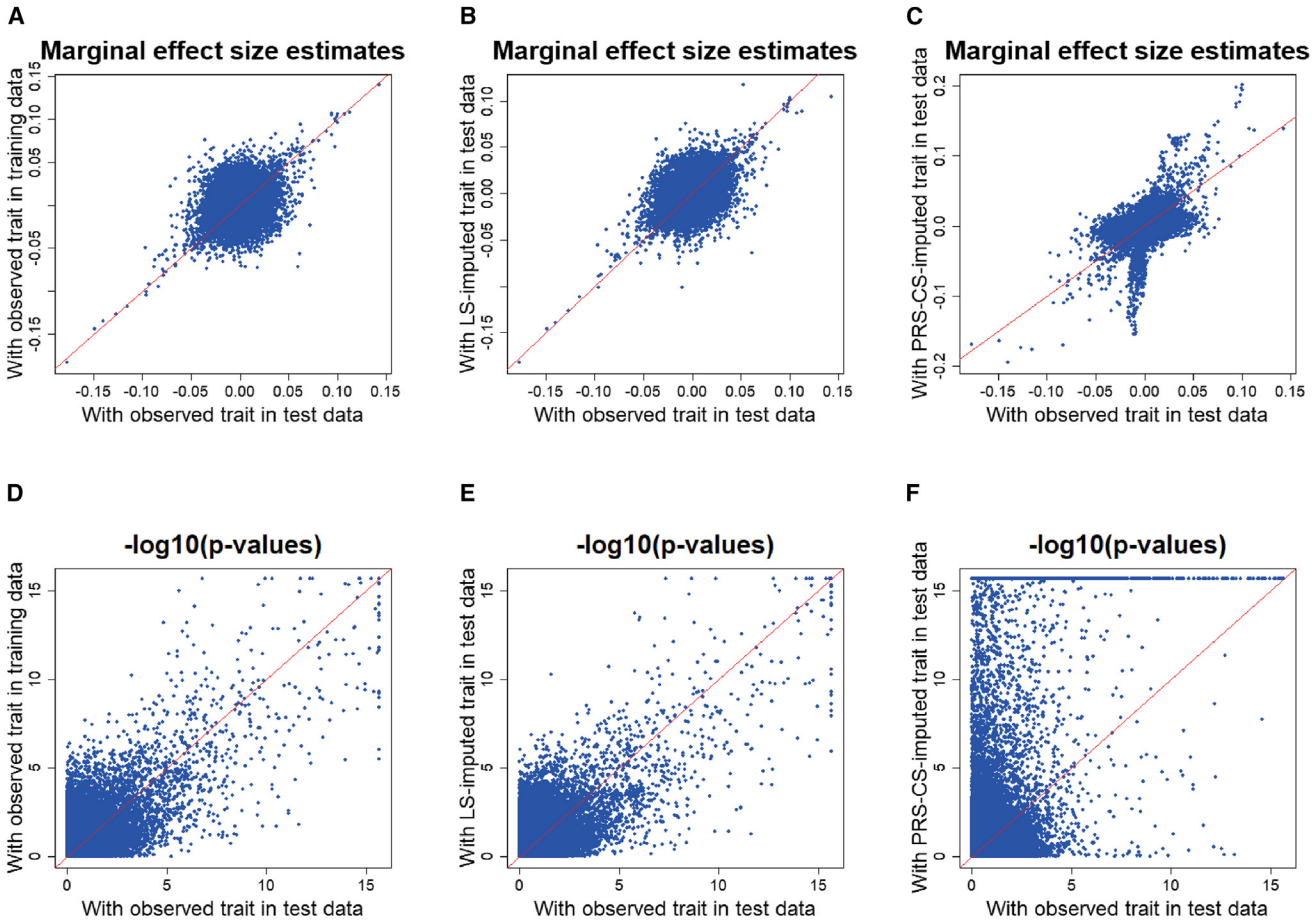

Figure 7. Comparison of the estimated marginal effect sizes (A–C) and –log10(p values) (D–F) under the recessive model: (A and D) calculated with the observed (HDL) trait (and genotypes) in the training and test data respectively; (B and E) with the observed (HDL) trait in the test data and LS-imputed trait in the test data; (C and F) with the observed (HDL) trait in the test data and PRS-CS-imputed trait in the test data

We reached the same conclusion under the dominant model (Figure S12) and recessive model (Figure 5) respectively: using the trait values imputed by our LS-imputation method, the GWAS results for the test data were in general quite similar to, though a bit more conservative than, those based on the observed trait values. In contrast, using PRS-CS returned too many significant associations.

We can further compare the estimated marginal effect sizes and their p values (for the 50,000 SNPs used in implementing our method) between various methods. As shown in Figures 6 (under the additive model) and 7 (under the recessive model), and in supplemental information S13 (under the dominant model), using the trait values imputed by our method, we could infer the marginal associations with the results similar to those achieved with the use of the observed trait values under each of the three genetic models. For comparison, again our method performed much better than PRS-CS for the purpose of this analysis.

As a summary, we show the Venn diagrams for the significant SNPs or loci identified under the various models in Figure 8. The significance cutoff was 5×10^{-8}. Each locus was defined as one of the 1,703 (approximately) independent LD blocks;[36] if a locus contained at least one significant SNP, it was declared to be a significant locus. It is clear that using the LS-imputed trait values gave the results in high agreements with those from using the observed trait values; in particular, the differences of the results between using the imputed and observed trait values were no more than that of using the observed trait values between the training and test data. As expected, the results from using the imputed trait values were similar to not only those of the test data with the observed trait values but also those of the training data, due to its use of both the training and test data.

SNP-SNP interactions

Here we consider whether we can detect SNP-SNP interactions based on imputed trait values compared with using the observed trait values. Based on the training data (and the additive model unless specified otherwise), we detected 1,758 marginally significant SNPs at $p < 10^{-6}$; after removing those in high LD (i.e., correlation > 0.99), we had 1,652 SNPs. We tested all pairwise interactions among the 1,652 SNPs; although the main effects of two SNPs in each pair were included in the model, we only tested the significance of their interaction term.

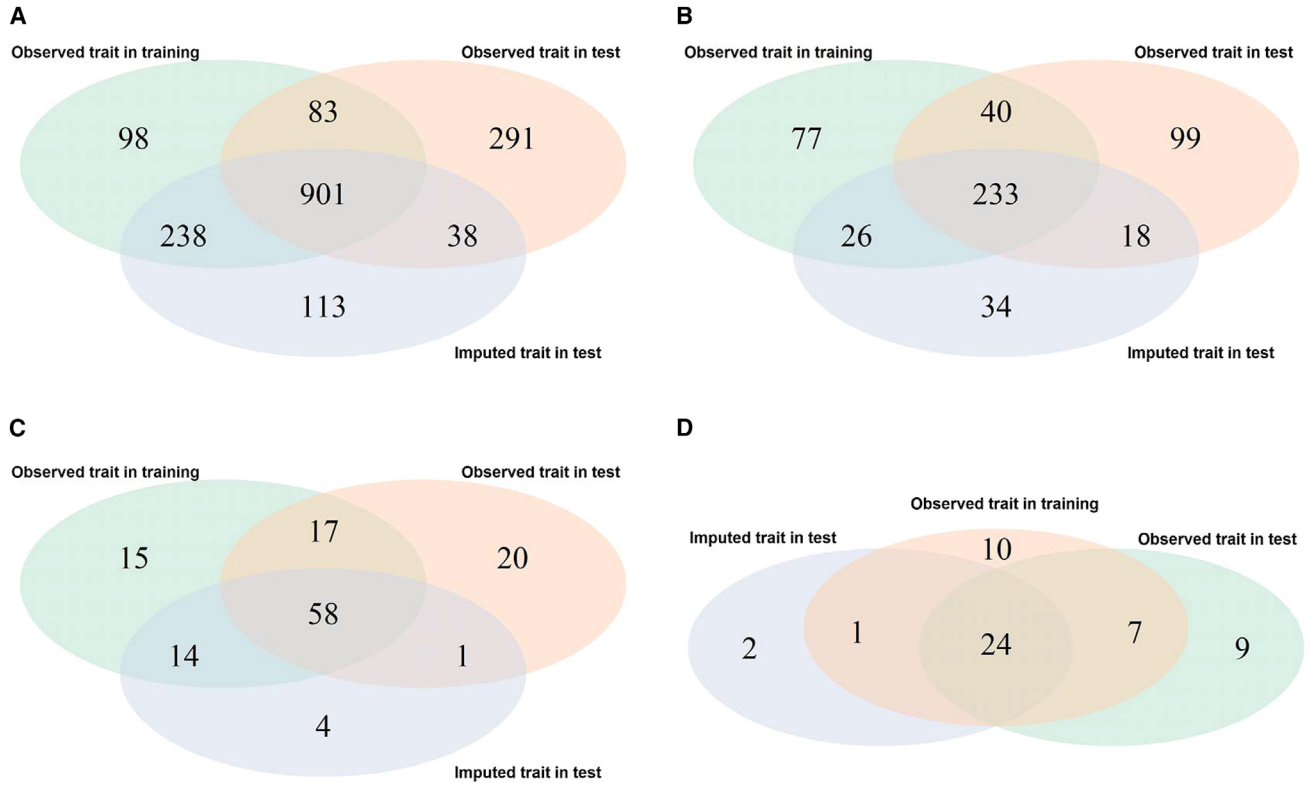

Figure 8. Venn diagrams for the significant SNPs (A and B) or loci (C and D) identified under the additive model (A and C) or under the recessive model (B and D)

Specifically, for any two of the 1,652 SNPs, we fitted a linear regression model:

$$Y_i = \alpha_0 + \mathrm{SNP}_{1i} \times \alpha_1 + \mathrm{SNP}_{2i} \times \alpha_2 + \mathrm{SNP}_{1i} \times \mathrm{SNP}_{2i} \times \alpha_{12} + e_i,$$

where Y_i, SNP_{1i}, and SNP_{2i} were the observed (or imputed) trait value and the two SNPs for individual i, and e_i was the error term. We applied the Wald test on the null hypothesis H_0: $\alpha_{12} = 0$.

Figure 9 compares the SNP-SNP interaction estimates and their p values for the test data using the observed trait values with using (A&D) the observed trait values (and genotypes) in the training data and (B&E) LS-imputed and (C&F) PRS-CS-imputed trait values in the test data. It is clear that using the LS-imputed trait values gave the results quite similar to those using the observed trait values, though the former might give slightly more conservative p values. Compared with PRS-CS, the LS-imputation method performed much better for the purpose of estimating and testing SNP-SNP interactions.

Figure 10 summarizes the results in terms of the significant SNP-SNP interactions and locus-locus interactions. Since we first searched the genome for marginally significant SNPs before testing them pairwise, we used the Bonferroni adjustment to obtain the significance cutoff of 2.5×10^{-8}, and as before, each locus was defined as one of the 1,703 (approximately) independent LD blocks.[36] For any significant SNP pair, we identified the locus of each of its SNPs and thus a significant locus-locus pair. There was a high agreement between the results of using the LS-imputed trait values and those from using the observed trait values in either the training or test data.

Trait prediction with a nonlinear model

We compared the trait prediction performance using the nonlinear RF model trained with either the observed or imputed (HDL) trait values. Specifically, we considered the 1,652 SNPs marginally significantly associated with the trait based on the training data (as in the previous section for SNP-SNP interaction detection), along with either the observed or imputed trait values in a random subset of 70% test data to train an RF model. Then we used the remaining 30% of the test data as the validation data to compare the predicted trait values from various RF models. Our goal was to assess the extent of the agreement between the predicted trait values from the RF models trained with either the observed or imputed trait values. As shown in Figure 11, the correlation of the predicted trait values between using the observed and LS-imputed trait values (at 0.722) was slightly higher than that (0.658) between the observed and PRS-CS-imputed ones, suggesting that our LS-imputed trait values retained more information about SNP-trait associations, possibly nonlinear, in the original data than that of the PRS-CS-imputed ones.

Other results

In the supplemental information S3.3, we compared the performance of the LS-imputation method with various

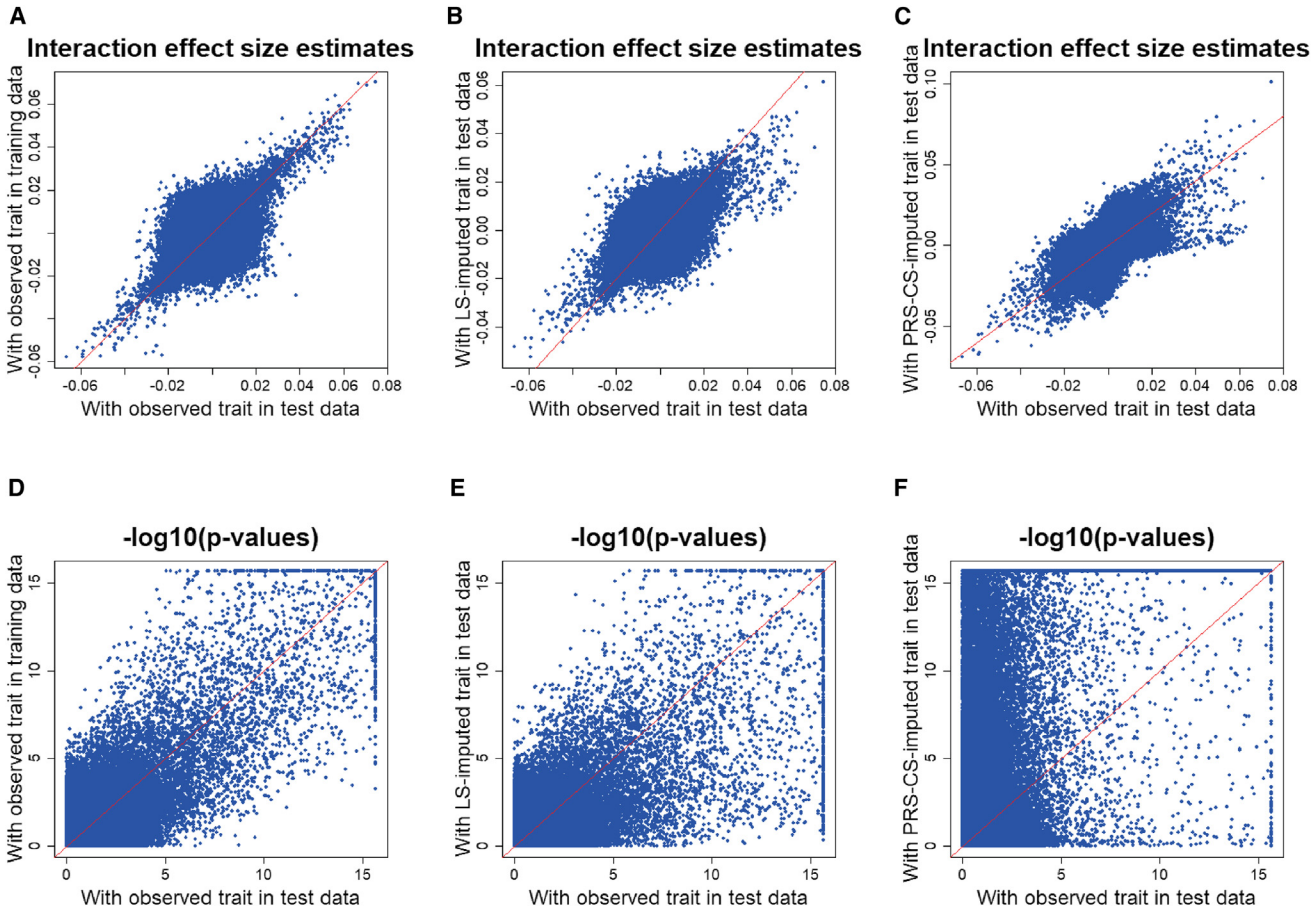

Figure 9. Comparison of the estimated SNP-SNP interaction effect sizes (A–C) and their –log10(p values) (D–F): (A and D) calculated with the observed (HDL) trait values (and genotypes) in the training and test data respectively; (B and E) with the observed and LS-imputed trait values in the test data; (C and F) with the observed and PRS-CS-imputed trait values in the test data

values of the training sample size (n_1), test sample size (n_2), SNP number (p), and batch size (m). As expected, usually the larger n_1 and p, the better is the performance; the results were not sensitive to n_2 as long as $n_2 \geq 25,000$. On the other hand, the batch size m also mattered in a complicated way: it was not necessarily better to use a larger (or smaller) batch size. Note that m corresponds to n_2 in our analysis in statistical properties of LS-imputation, where it is indicated that its optimal choice cannot be too large or too small. Our current solution is to choose an m giving the (additive) marginal analysis results (i.e., in terms of the marginal association estimates and their SEs) with imputed traits in the test data similar to those from the training data. Note that, if the sample sizes from the training and test data are different, we need to rescale the SEs from one of the two samples to make them comparable. For any given SNP with its SEs as SE1 and SE2 from the training and test data (of sample size n_1 and n_2), respectively, we would rescale the SE from the test data as $\sqrt{n_2/n_1}SE2$.

Furthermore, as shown in supplemental information S3.5 and S3.6, we have also compared the computational speed and stability of several implementations of the LS-imputation method, In the end, we found that inverting a regularized XX' (i.e., $XX' + \lambda I$), denoted inv ($\lambda = 10^{-6}$), was always fast and stable, so it was chosen as the default implementation. Second, as shown in supplemental information S3.4, we found that the results from the WLS were quite similar to those from the default ordinary least-squares (OLS) method. Third, we applied the LS-imputation method to a smaller subset of the UKB GWAS data of trait BMI, obtaining similar results.

Simulations

As shown in supplemental information S4, we did a simulation study mimicking the real UKB HDL GWAS data, confirming the good performance of our method.

Binary traits

As shown in supplemental information S2, we have extended the LS-imputation method to binary traits and applied it to the UKB hypertension GWAS data, obtaining promising results.

Discussion

We have proposed a nonparametric method for large-scale imputation of (the genetic components of) a trait based

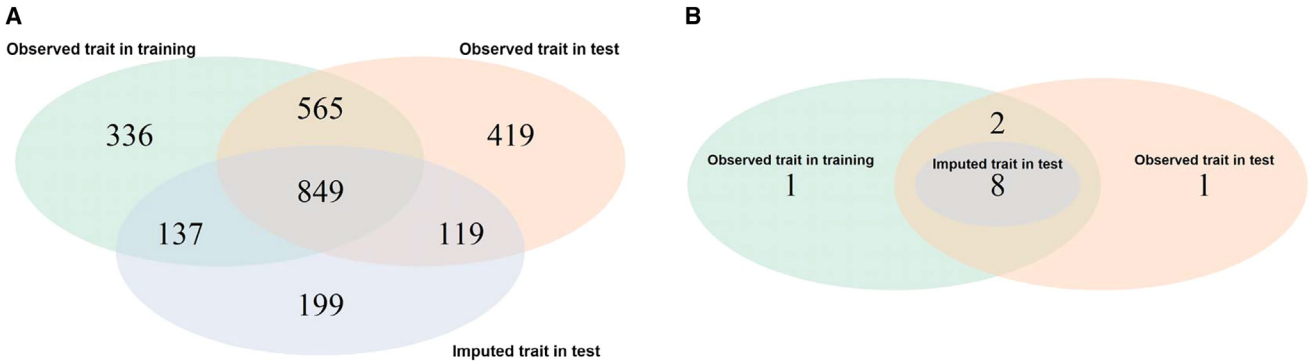

Figure 10. Venn diagrams for the significant (A) SNP-SNP interactions and (B) locus-locus interactions identified

on a GWAS summary dataset and a large dataset of individual-level (genome-wide) genotypes. We emphasize that, although linear marginal association estimates are used for trait imputation, we impose no assumption on the specific functional form of the genetic component of the trait, so the imputed trait values can be used for both linear and nonlinear SNP-trait association or prediction analysis. This may sound surprising, and we offer some intuitive explanations from two aspects. First, as shown theoretically, if the test genotypic data are the same as the training genotypic data, we will recover the trait values of the training data exactly, which clearly contain information about possible SNP-trait nonlinear associations. Second, more generally, for any given test sample, its imputed trait value is a linear combination of the trait values in the training data; the weights in the linear combination depend on the similarities between the test genotype and the training genotypes. Again, a linear combination of the trait values is expected to contain information about possible nonlinear SNP-trait associations. Another distinct feature of the proposed method is its borrowing information across the genotypes of the individuals whose trait values are to be imputed: if some individuals are similar to each other (in terms of their genotypes), their imputed trait values would incorporate each other's genotypes (in addition to their own). Hence, plus its nonparametric nature, our method is most suitable for large-scale trait imputation simultaneously for a large set of individuals.

Compared with a leading PRS method, PRS-CS, which is linear model-based as all existing PRS methods for summary data, our LS-imputation method performed much better in imputing the trait for subsequent linear or nonlinear SNP-trait association analyses, such as in estimating additive or non-additive effects of SNPs, and in detecting SNP-SNP interactions. We emphasize that in general any existing PRS method is not designed for trait imputation for subsequent association analyses as targeted here; however, since the application of a PRS method seems to offer an alternative, we assessed and compared the performance of PRS-CS for our problems here. In fact, in our UKB HDL GWAS data example, the linear model-based PRS-CS did better than the nonparametric LS-imputation method in giving predicted/imputed trait values more highly correlated with the observed trait values, perhaps due to that linear/additive effects of SNPs dominated the heritability.[34,35] However, since PRS-CS, as many other model-based PRS methods, assumes linear effects of some *specific* SNPs on a trait, its imputed trait values are based on the *estimated* linear effects of the selected SNPs (which may not be truly causal or associated ones) and accordingly are not suitable for subsequent association analyses, though they are useful for prediction. Relatedly, although other methods have been proposed to impute a few missing values of a focal trait using other traits,[37–39] they are not suitable for our purpose of large-scale trait imputation for downstream genetic association analysis because of the loss of specificity: by definition, any genetic variants associated with a trait used to impute the focal trait are expected to be associated with the *imputed* focal trait, even not truly associated with the (observed) focal trait.

Our methods require $p > n_2$ (or more generally, $p > m$ if batches are used); that is, the number of the SNPs chosen to be used for trait imputation is larger than the test (or batch) sample size. The basic idea of our method is that each marginal SNP-trait association estimate in the GWAS summary data imposes a constraint on the possible values of the trait; with $p > n_2$, we have more constraints than the number of unknown trait values, thus uniquely determining the trait values. Under this condition (and that of no closely related individuals in the test data), the non-full-rank of the genotype matrix and thus the use of the Moore-Penrose inverse are completely due to the effects of centering the genotype matrix (to account for unknown intercepts), leading to the effects of centering the imputed trait values. We expect that other generalized inverses may be used (but with possibly different properties of the imputed trait values).

The current implementation of our proposed method can be further extended. First, we have only considered using marginal associations from common variants. Although rare variants can be equally used, they are expected to contain less association or heritability information than common variants and to have lower genotyping quality in array-based GWAS data. As more large-scale sequencing data become available and computing power keeps going up, it will be worthwhile to explore the use of rare variants

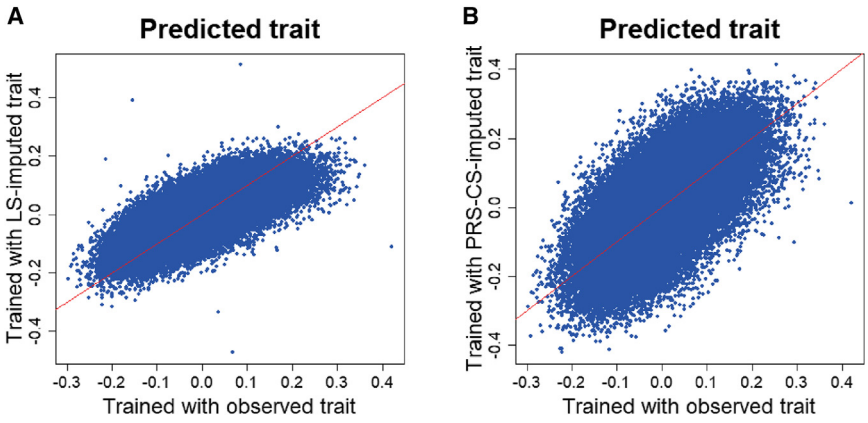

Figure 11. Comparison of the predicted (HDL) trait values with random forests models: (A) trained with the observed versus LS-imputed trait values; (B) with the observed versus PRS-CS-imputed ones

for trait imputation. Second, we note that, to save computing cost (and to simplify subsequent analyses), we currently have ignored the correlations and unequal variances of the imputed trait values, leading to possibly slightly more conservative (or liberal) inference; accounting for the correlations is straightforward in theory but requiring much more and even unrealistic computing resources with a likely return of only minor performance improvement, so we decided not to pursue it, though it may be further explored in the future. Third, we have considered some of the perhaps more extreme and challenging problems of statistical inference using only imputed traits; other more practical applications include using the imputed traits to augment a complete individual-level GWAS dataset for inference or prediction or using the imputed data to generate prior information or as partial validation data for other GWAS analyses. Fourth, we expect that more efficient algorithms to handle larger data will be useful and needed. Instead of using the OLS to impute the (genetic components of) trait values, using generalized least-squares (to account for correlated marginal association estimates with varying variances) may offer statistically more efficient imputation, though it will be computationally even more demanding (mainly in dealing with a large covariance matrix of $\widehat{\beta}^*$). Fifth, although we have extended the method to binary traits, further evaluations and applications are warranted. Finally and more importantly, it would be worth exploring other applications with imputed traits beyond those showcased here. With the ever-increasing availability of GWAS summary data of various traits and the emergence of large-scale biobanks, the proposed method can be applied to impute traits that are not measured or not (fully) available in a biobank (e.g., status of a late-onset disease not yet fully manifesting in a cohort of younger individuals), which can be then used for analyses (with suitable adjustments) along with other available traits and genotypes. As an example, while individual-level GWAS data are required to fit a neural network to detect genes with possibly nonlinear effects of gene expression on Alzheimer disease (AD) in transcriptome-wide association studies,[40] the sample size of such data is small; on the other hand, we have large-scale AD GWAS summary data[41] and UKB individual-level genotypes available, but due to the late-onset nature of AD, we do not have many AD cases in the UKB data. Applying our method to impute the AD status for genotyped individuals in UKB would largely augment the sample size of individual-level AD GWAS data, thus improving model fitting and subsequent statistical power. This application motivated the development of our LS-imputation method and is currently under investigation.

Data and code availability

One needs to apply to UK Biobank for approval to access the individual-level GWAS data used here. The code generated during this study is available at https://github.com/ren328/LSimputing.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2023.100197.

Acknowledgments

We thank the reviewers for many helpful and constructive comments. This research was supported by NIH grants U01 AG073079, R01 AG065636, R01 AG069895, RF1 AG067924, R01 HL116720, and R01 GM126002 and by the Minnesota Supercomputing Institute at the University of Minnesota. The application number to access the UK Biobank data is #35107.

Declaration of interests

The authors declare no competing interests.

Received: December 22, 2022
Accepted: April 7, 2023

Web resources

LS-imputation software: https://github.com/ren328/LSimputing.

PRS-CS software: https://github.com/getian107/PRScs.

UK Biobank data: https://www.ukbiobank.ac.uk/.

References

1. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. Nucleic Acids Res. *47*, D1005–D1012.
2. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
3. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Weedon, M.N., Loos, R.J., Genetic investigation of ANthropometric Traits (GIANT) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM), Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375.
4. Ma, Y., and Zhou, X. (2021). Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. *37*, 995–1011.
5. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504.
6. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
7. Speed, D., and Balding, D.J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat. Genet. *51*, 277–284.
8. Song, S., Jiang, W., Zhang, Y., Hou, L., and Zhao, H. (2022). Leveraging LD eigenvalue regression to improve the estimation of SNP heritability and confounding inflation. Am. J. Hum. Genet. *109*, 802–811.
9. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Patterson, N., Robinson, E.B., ReproGen Consortium, Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., Perry, J.R.B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
10. Zhang, Y., Lu, Q., Ye, Y., Huang, K., Liu, W., Wu, Y., Zhong, X., Li, B., Yu, Z., Travers, B.G., et al. (2021). SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biol. *22*, 262.
11. Burgess, S., Davey Smith, G., Davies, N.M., Dudbridge, F., Gill, D., Glymour, M.M., Hartwig, F.P., Holmes, M.V., Minelli, C., and Relton, C.L. (2020). Guidelines for performing Mendelian randomization investigations. Wellcome Open Res. *4*, 186.
12. Zuber, V., Grinberg, N.F., Gill, D., Manipur, I., Slob, E.A.W., Patel, A., Wallace, C., and Burgess, S. (2022). Combining evidence from Mendelian randomization and colocalization: review and comparison of approaches. Am. J. Hum. Genet. *109*, 767–782.
13. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., Cox, N.J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
14. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
15. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.
16. Holzinger, E.R., Verma, S.S., Moore, C.B., Hall, M., De, R., Gilbert-Diamond, D., Lanktree, M.B., Pankratz, N., Amuzu, A., Burt, A., et al. (2017). Discovery and replication of SNP-SNP interactions for quantitative lipid traits in over 60,000 individuals. BioData Min. *10*, 25.
17. Zhou, J., Passero, K., Palmiero, N.E., Müller-Myhsok, B., Kleber, M.E., Maerz, W., and Hall, M.A. (2020). Investigation of gene-gene interactions in cardiac traits and serum fatty acid levels in the LURIC Health Study. PLoS One *15*, e0238304.
18. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.
19. Fryett, J.J., Morris, A.P., and Cordell, H.J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. Genet. Epidemiol. *44*, 425–441.
20. Grinberg, N.F., and Wallace, C. (2021). Multi-tissue transcriptome-wide association studies. Genet. Epidemiol. *45*, 324–337.
21. Okoro, P.C., Schubert, R., Guo, X., Johnson, W.C., Rotter, J.I., Hoeschele, I., Liu, Y., Im, H.K., Luke, A., Dugas, L.R., et al. (2021). Transcriptome prediction performance across machine learning models and diverse ancestries. HGG Adv. *2*, 100019.
22. Ma, W., Lau, Y.L., Yang, W., and Wang, Y.F. (2022). Random forests algorithm boosts genetic risk prediction of systemic lupus erythematosus. Front. Genet. *13*, 902793.
23. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.
24. Ghose, U., Sproviero, W., Winchester, L., Fernandes, M., Newby, D., Ulm, B.S., Shi, L., Liu, Q., Adams, C., Albukhari, A., et al. (2022). Genome wide association neural networks (GWANN) identify novel genes linked to family history of Alzheimer's disease in the UK Biobank. Preprint at medRxiv. https://doi.org/10.1101/2022.08.15.503991.
25. Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S., Puiggròs, M., Salvoro, C., Miguel-Escalada, I., Carey, C.E., Cole, J.B., Rüeger, S., Atkinson, E., et al. (2021). The impact of non-additive genetic associations on age-related complex diseases. Nat. Commun. *12*, 2436.
26. O'Connor, M.J., Schroeder, P., Huerta-Chagoya, A., Cortés-Sánchez, P., Bonàs-Guarch, S., Guindo-Martínez, M., Cole, J.B., Kaur, V., Torrents, D., Veerapen, K., et al. (2022). Recessive genome-wide meta-analysis illuminates genetic architecture of type 2 diabetes. Diabetes *71*, 554–565.
27. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.
28. DasGupta, A. (2008). Central Limit theorems for dependent sequences. In Asymptotic Theory of Statistics and Probability (Springer Texts in Statistics. Springer). https://doi.org/10.1007/978-0-387-75971-5_9.
29. Chafai, D. (2009). Singular Values of Random Matrices. https://djalil.chafai.net/docs/sing.pdf.
30. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
31. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell,

J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
32. Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., et al. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. PLoS Genet. *17*, e1009021.
33. Zhou, G., and Zhao, H. (2021). A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet. *17*, e1009697.
34. Hivert, V., Sidorenko, J., Rohart, F., Goddard, M.E., Yang, J., Wray, N.R., Yengo, L., and Visscher, P.M. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. Am. J. Hum. Genet. *108*, 786–798.
35. Pazokitoroudi, A., Chiu, A.M., Burch, K.S., Pasaniuc, B., and Sankararaman, S. (2021). Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. Am. J. Hum. Genet. *108*, 799–808.
36. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.
37. Dahl, A., Iotchkova, V., Baud, A., Johansson, Å., Gyllensten, U., Soranzo, N., Mott, R., Kranis, A., and Marchini, J. (2016). A multiple-phenotype imputation method for genetic studies. Nat. Genet. *48*, 466–472.
38. Hormozdiari, F., Kang, E.Y., Bilow, M., Ben-David, E., Vulpe, C., McLachlan, S., Lusis, A.J., Han, B., and Eskin, E. (2016). Imputing phenotypes for genome-wide association studies. Am. J. Hum. Genet. *99*, 89–103.
39. An, U., Pazokitoroudi, A., Alvarez, M., Huang, L.Y., Bacanu, S., Schork, A.J., Kendler, K., Pajukanta, P., Flint, J., Zaitelen, N., et al. (2022). Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.15.503991.
40. He, R., Liu, M., Lin, Z., Zhuang, Z., Shen, X., and Pan, W. (2023). DeLIVR: a deep learning approach to IV regression for testing nonlinear causal effects in transcriptome-wide association studies. Biostatistics, kxac051.
41. Bellenguez, C., Küçükali, F., Jansen, I.E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A.C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., et al. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat. Genet. *54*, 412–436.